ELSEVIER

Computer Networks and ISDN Systems 30 (1998) 1327–1340

COMPUTER NETWORKS and ISDN SYSTEMS

Standardization of traffic measurements and models for broadband networks: open issues

Arthur W. Berger [a,*], Maurizio Naldi [b,1], Livia De Giovanni [c,2], Manuel Villén-Altamirano [d,3]

[a] *Lucent Technologies, 101 Crawfords Corner Rd., Holmdel, NJ 07733, USA*
[b] *Telecom Italia, Via Valcannuta 250, 00166 Roma, Italy*
[c] *Università del Molise, Via F. De Sanctis, 86100 Campobasso, Italy*
[d] *Telefónica I + D, Emilio Vargas 6, 28043 Madrid, Spain*

Accepted 14 March 1998

Abstract

The International Telecommunications Union (ITU) Study Group 2 is preparing recommendations concerning traffic engineering to support broadband services, for which measurements and modeling are a key component. The present paper reviews three problem areas that could be of interest to teletraffic researchers and for which the ITU would appreciate future contributions. The first area concerns the potential interactions of standardized, algorithmic traffic descriptors with stochastic models for cell traffic variables. The second area is on the appropriate definition of cell-level and call-level measurements that enable the inference of important characteristics of the traffic and of the performance of the network. The last area concerns long-range dependent versus Markovian models and the need for future work to consider the endogenous character of data traffic given feedback controls. Crown Copyright © 1998 Published by Elsevier Science B.V. All rights reserved.

Keywords: B-ISDN; ATM; ITU; Standardization; Traffic measurements; Traffic models; Traffic engineering

1. Introduction

The International Telecommunications Union (ITU), a specialized agency of the United Nations, is preparing recommendations concerning traffic engineering to support Broadband-Integrated Services Digital Networks (B-ISDNs), for which measurements and modeling are a key component. B-ISDNs are communication networks that support a variety of services, including voice, video and data, on a common network whose transmission links are typically in the range of 34 Mega-bits per second (Mbps) to 2.4 Giga-bits per second (though some access links may be as slow as 1.5 or 2 Mbps). Within the ITU and much of the telecommunications industry, B-ISDNs will be implemented using the technology of Asynchronous Transfer Mode (ATM), which is a packet network with small, fixed-size packets called cells. ATM networks are currently being deployed, primarily, as an infrastructure in public and private

[*] Corresponding author. E-mail: awberger@lucent.com.
[1] E-mail: m.naldi@ieee.org.
[2] E-mail: lidegio@tin.it.
[3] E-mail: manolo@tid.es.

PII: S0169-7552(98)00143-3

networks. A public telecommunications service provider can use the ATM infrastructure to internally integrate its various service offerings over a common backbone network, thereby realizing savings in operations, administration and capital costs. Also, public carriers are offering ATM-based services that businesses, government agencies, and universities can in turn use to obtain their own ATM infrastructure to support their private networking needs.

The present paper reviews three problem areas regarding measurements and modeling to support traffic engineering for B-ISDNs:

1. Traffic descriptors and traffic variables,
2. Cell and call-level measurements,
3. Long-range dependent and Markovian traffic models.

For each topic, background material is provided followed by the exposition of open issues of concern to the ITU. The topics were selected for their potential interest to teletraffic researchers, and contributions on them will be appreciated by the ITU.

1.1. The standardization panorama

In the ITU-T, traffic engineering has traditionally fallen under the responsibility of Study Group 2 (Network and Service Operation), see Fig. 1. In contrast, Study Group 13 (General Network Aspects) is responsible for initial studies on B-ISDN including the definition of network architectures, capabilities and interfaces, and network performance.

In this context, Study Group 13 has developed Recommendation I.371 [1], which defines the traffic contract and the different traffic controls for B-ISDN. For network performance, it has developed Recommendation I.356 [2] for cell level performance and I.358 [3] for call level performance. Some of the performance parameters defined in the recommendations are traffic related.

Study Group 2 is developing recommendations on traffic engineering for B-ISDN within the framework of the Study Group 13 recommendations. Study Group 2 recommendations specifically devoted to B-ISDN are E.716 [4] on traffic modeling, and E.735 [5], E.736 [6] and E.737 [7] on traffic control and dimensioning, and presently under development E.726 [8] on traffic related performance parameters and E.745 [9] on cell level traffic measurements. Other related Study Group 2 recommendations are E.500 [10] (which is presently under revision) and E.493 [11] which cover call-level traffic measurements in several networks including B-ISDN.

2. Traffic descriptors and traffic variables

The various characterizations of the cell flow on an ATM connection can be grouped into two types: (1) standardized traffic descriptors and (2) stochastic

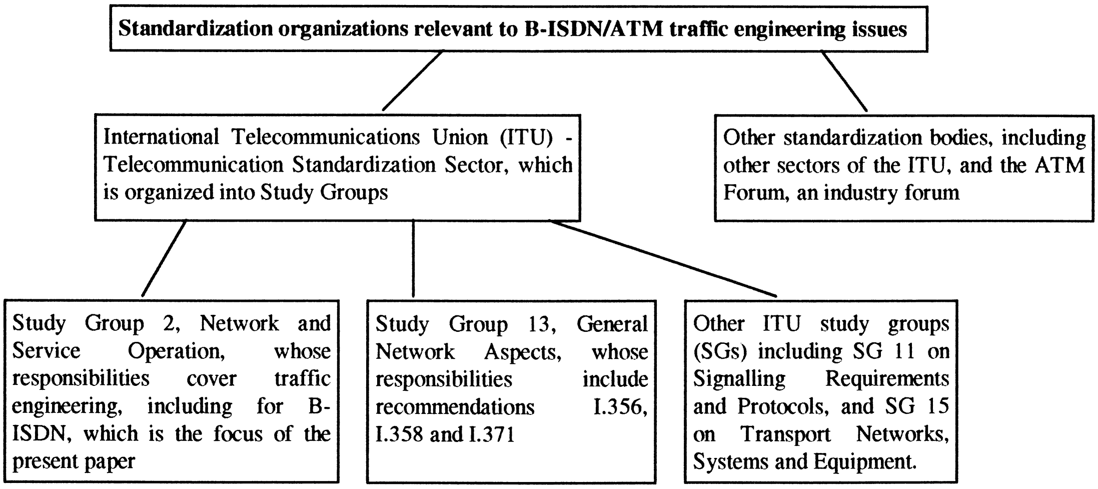

Standardization organizations relevant to B-ISDN/ATM traffic engineering issues
International Telecommunications Union (ITU) - Telecommunication Standardization Sector, which is organized into Study Groups
Other standardization bodies, including other sectors of the ITU, and the ATM Forum, an industry forum
Study Group 2, Network and Service Operation, whose responsibilities cover traffic engineering, including for B-ISDN, which is the focus of the present paper
Study Group 13, General Network Aspects, whose responsibilities include recommendations I.356, I.358 and I.371
Other ITU study groups (SGs) including SG 11 on Signalling Requirements and Protocols, and SG 15 on Transport Networks, Systems and Equipment.

Fig. 1. Standardization organizations relevant to B-ISDN traffic engineering issues.

cell traffic variables. Standardized traffic descriptors are defined in terms of an algorithm that determines whether or not each arriving cell is conforming to the traffic descriptor. In contrast, cell traffic variables are parameters that are used in a stochastic model of the cell flow.

2.1. Review

2.1.1. Terminology on standardized traffic descriptors

As stated in ITU Recommendation I.371 [1], the traffic descriptor parameters of a connection are the set of traffic parameters used during the connection set-up to capture the traffic characteristics of the connection at a standardized interface. The traffic descriptor parameters of a connection at the user-network interface (UNI) are declared by the user and surveyed by the network (policing function). Thus the traffic descriptor parameters, apart from providing information useful for performing resource allocation, must satisfy the following conditions:

- be understandable for the user (or terminal), which has to be able to declare them prior to the realized cell flow; conformance should be possible;
- be enforceable by the policing function at the UNI and inter-network interfaces (INI).

The above two conditions imply that the traffic descriptor parameters should be simple and they should be defined algorithmically, and thus form deterministic bounds on the traffic characteristics of the connection. Since the traffic descriptor parameters of a connection are a part of the contract between the user and the network operator, they have been standardized. To date the traffic parameters defined in I.371 are the peak cell rate (PCR) and the pair formed by the Sustainable Cell Rate (SCR) and Intrinsic Burst Tolerance (IBT). The SCR/IBT are defined in terms of the leaky-bucket algorithm (a.k.a. the virtual scheduling algorithm or the generic cell rate algorithm). At a standardized interface, the UNI or INI, the PCR is associated with a cell-delay-variation tolerance (CDVT) to account for the delay variation that may occur upstream from the interface. The pair PCR/CDVT is also defined in terms of the leaky-bucket algorithm, and it determines an operational bound on the peak cell rate of the connection at each interface. Similarly at a given interface, the SCR is an upper bound on the realizable average rate of a connection, and the SCR/IBT (with a tolerance added to the IBT to account for upstream cell delay variation) jointly determine the maximum conforming burst sent at the PCR. The parameters PCR, SCR, and IBT constitute intrinsic traffic characteristics of the connection at the source and their value is unchanged at all interfaces of the connection. In contrast, the cell-delay-variation tolerances may be different at each interface.

2.1.2. Cell traffic variables

Cell traffic variables are used in stochastic models of ATM traffic; they are typically moments or quantiles or other parameters pertaining to the distribution of random variables that model the cell arrival process. Stochastic models have received tremendous attention in the literature. In the context of ITU recommendations, cell traffic variables are discussed in Recommendation E.716 [4], where a survey of the most usual ones is presented. In particular, four alternative approaches for defining cell traffic variables are presented in E.716, depending on model viewpoint:

- the burst structure of the cell flow,
- the number of cell arrivals in given intervals,
- the interarrival times,
- the number of cells exceeding certain rates.

The traffic engineer has the task of selecting the most significant traffic variables, from the viewpoint of determining the impact of the connection on the network performance.

2.1.3. Service-type designator

The service-type designator is a qualitative traffic descriptor also specified in Recommendation I.371 [1], which provides an implicit declaration of the traffic characteristics of a connection. To date, this designator has received little attention, and no particular service-types have been defined. Examples might be video-teleconferencing with a particular coding standard, or fax of a particular generation. The end-system would just need to specify which service-type pertains for the given connection, and the network operator would know a good deal about the characteristics of the connection, based on prior knowledge. This knowledge could be in terms of both standardized traffic descriptors (e.g. the application would never generate cells faster than a given rate) and in terms of stochastic cell traffic variables, as well as quality of service (QoS) requirements that would need to be satisfied. The Service-Type designator could also be used in conjunction with an ATM traffic parameter such as the PCR.

2.1.4. Discussion

Superficially, the standardized traffic descriptor is similar to the cell traffic variables as they both describe the traffic in some sense. However, there are fundamental differences. The parameter values of the traffic descriptor pertain to the given realization of the cell flow. Both the source and the network should be able to determine throughout the duration of the connection whether each cell arrival is conforming. In contrast, of course, a given realization of a finite cell flow may deviate markedly from the values of the cell traffic variables. Standardized traffic descriptors provide a deterministic bound on the cell flow, while cell traffic variables provide information on the expected characteristics of the cell flow and allow for a more thorough characterization of the traffic.

2.2. Open issues

The following open issues concern the interplay between traffic variables and traffic descriptors.

1. Given that traffic engineering models are generally based on the use of traffic variables, what traffic variables should be used to describe individual connections for purposes of connection admission control (CAC) and dimensioning?
   - 1.1. Should we use cell traffic variables corresponding to the ‘‘worst case’’ traffic characteristics compatible with declared values of the traffic descriptor parameters? This has the disadvantage that the network may be over-dimensioned.
   - 1.2. Should we use cell traffic variables estimated from measurements or from the knowledge of the service type (or, for estimating the cell-delay-variation, from assumptions on the upstream networks)? This has the disadvantage that it could require significant resources for measurements, and the network operator is less certain of providing the committed quality of service to the admitted connections.
2. In case 1.2 above, how should the traffic variables incorporate, if at all, the standardized traffic descriptors that pertain at given interfaces?
   - 2.1. For example, in the context of public ATM networks and variable-bit-rate (VBR) video being carried over a real-time VBR ATM connection, the stochastic model of the cell flow should incorporate the influence of the standardized traffic descriptors PCR and SCR/IBT.
   - 2.2. Likewise, for a data application that is being supported by a non-real-time VBR ATM connection (which is analogous to a frame-relay connection), the stochastic model of the cell flow from the data application should also incorporate the influence of the standardized traffic descriptors PCR and SCR/IBT.
3. Would the definition of additional standardized traffic descriptors yield a substantial benefit? For what services?
   - 3.1. Can cell traffic variables from stochastic models provide insight into the judicious definition of standardized traffic descriptors?
4. Is the Service-Type designator useful?
   - 4.1. For a given application, what are particular definitions of Service-Type, which could include both parameter values for the standardized traffic descriptors and the model of stochastic traffic variables, as well as QoS requirements?

3. Measurements

The growing demand for broadband services is accelerating the implementation of an operating network platform based on the ATM technique. Aside from all the procedures needed to set-up and run such a network, the traffic monitoring function cannot be overlooked. Of extreme interest is therefore the definition of a set of traffic measurements to be routinely performed on the ATM network, regarding traffic characterization, performance monitoring, and the monitoring of traffic-control actions. With the introduction of ATM, we now need measurements at the cell level, in addition to the measurements at the call level that were required of a circuit-switched network. Given the large variety of traffic sources acting as input to the network, the task is quite difficult, and the effort is still at an early stage.

In the following, an attempt is made to analyze the current status of the traffic measurement activity and to expose some of the problems still facing the network operators.

3.1. Contexts of application of traffic measurements

Traffic measurements are of interest to a large and diverse set of people, including the network planners and designers (who need long-term statistics to dimension tomorrow's networks), the Operations and Management Department (which is in charge of detecting sudden traffic surges, taking the appropriate Network Management countermeasures, and in addition collecting the statistics of interest to all the other users), and the Administration Department (which has to run the billing procedures). The simple collection of the desires of all the measurement users would soon lead to a large and unmanageable measurement book. The introduction of any new measurement represents in fact an additional burden to the processing capabilities of the ATM exchange, whose ultimate task is that of switching rather than measuring.

The first step of the definition process is nevertheless the categorization of measurement users and an understanding of their needs. To this purpose the following rough classification of the application contexts can therefore be useful (billing has not been included since it is generally not under the responsibility of traffic engineers):

- planning and dimensioning,
- performance monitoring,
- traffic control (including what is called Network Traffic Management in the telephone network).

These three activities are listed in order of growing time-sensitivity. As network planning is generally performed on a yearly basis, for dimensioning purposes traffic measurements can be processed off-line. On the contrary, traffic control procedures must be executed on a much tighter time scale: for example connection admission control (CAC) could use real-time measurements of occupancy that have been taken over a time scale of one minute.

Note that each of the applications above require both call level and cell level measurements.

3.2. Review

3.2.1. Measurements at the cell level

As was stated above, three types of measurements can be distinguished: those for traffic characterization, for performance monitoring and for monitoring of traffic-control actions. For traffic characterization, the cell traffic variables characterizing a cell flow from the point of view of its impact on network performance must be measured. ITU-T Recommendation E.716 [4] defines cell traffic variables and proposes four alternative sets of traffic variables for characterizing a cell flow. In draft Recommendation E.745 [9], work is being carried out on the feasibility of these measurements and on their attributes: the object, the frequency and the readout period. The object of the measurement could be an ATM connection, an ATM link, the connections of the same QoS class in an ATM link, etc. The frequency could be, for example, continuous (i.e. at every moment), or on a given periodic schedule, or on request for special studies. The readout period is the length of the time interval after which a measurement value is provided.

As to the variety of aims that can be served by cell-level measurements regarding traffic characterization, a first point to mention is the use of these measurements in the design of efficient ATM layer traffic controls algorithms, in particular CAC. In fact, aside from considering the traffic descriptor parameters declared for the new and currently established calls, the CAC algorithm may base its decisions on either past measurements (which allow to correlate the declared traffic descriptor parameters with cell traffic variables) or on real-time measurements of the currently established calls (Dynamic CAC). If the CAC is based on measurements, the measured cell traffic variables will influence the number of connections admitted, and hence the cell-level traffic measurements may be used in network dimensioning as well.

The framework of measurements for performance monitoring at the cell level is presented in the ITU-T Recommendation I.356 [2], prepared by Study Group 13, where the ATM layer Performance parameters are defined in terms of:

- *Measurement Points* (MPs),
- *Cell Reference Events* (CREs), i.e. events associated with cell transits through the measurement points, as the *cell exit* or the *cell entry event*,
- *Cell Transfer Outcomes*, defined through two corresponding cell transfer reference events at two different MPs (CRE1 at MP1 and CRE2 at MP2), such as *successful cell transfer outcome*, *lost cell outcome*, or *misinserted cell outcome* (see Fig. 2).

Using the above cell transfer outcomes, ATM cell transfer performance parameters (such as *cell loss ratio* or *cell transfer delay*) can be defined for two general Measurement Points.

After properly locating the two MPs within the network element (at the interfaces between layers in the ATM protocol stack according to recommendation I.356), the above framework is used by Study Group 2 in the ITU-T Recommendation E.745 [9] to define the performance measurements at the cell level that are required for traffic engineering.

A third group of measurements are those to monitor traffic-control actions. For example, the policing mechanism may reject cells which are non-conforming with the declared traffic descriptor. Although the operator is not responsible for the loss of non-conforming cells, it can be interested in monitoring the cells rejected by the policer to advice the users on an appropriate choice of the traffic descriptor parameters. The network operator is also interested in knowing the cells rejected by the policer at subsequent network interfaces. Another example of this

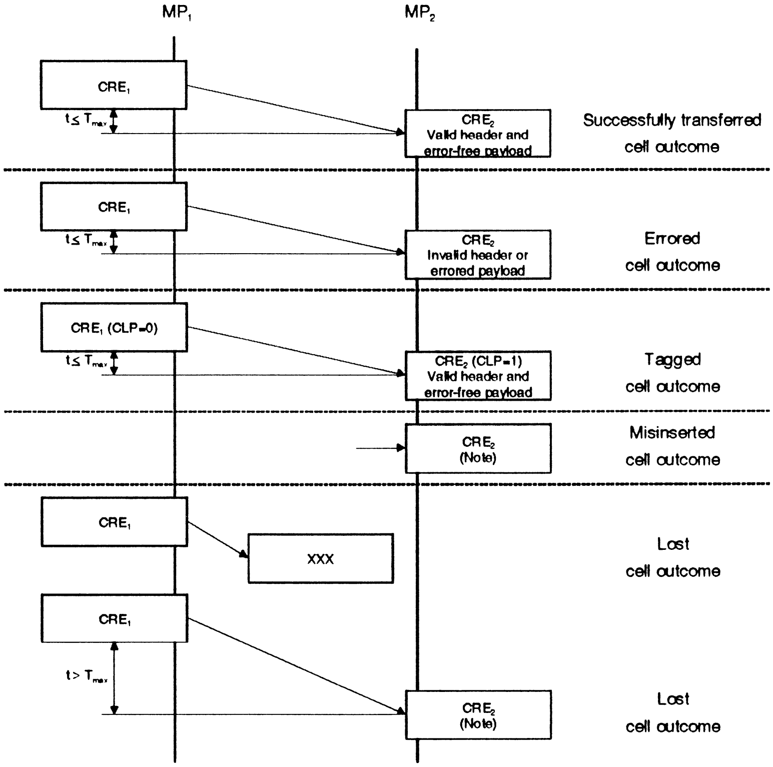


Fig. 2. Cell transfer outcomes (after Rec. I.356 [2]). t = time interval between cell reference events (CREs), $T_{\max}$ = upper limit on t in order that the cell not be judged as lost, CLP = cell-loss-priority bit in ATM cell header (CLP = 0 means a high priority cell, and CLP = 1 means low priority), XXX denotes a discarded cell. *Note*: Outcome occurs independent of cell content.

type measurement is the recording of the number of Resource Management cells sent within Available Bit Rate or ATM Block Transfer connections.

3.2.2. Measurements at the call level

The standardization activity devoted to ATM traffic measurements at the call level has been lower than at the cell level. It should be reminded that, at the call level, an ATM network has a number of similarities with a circuit-switched one. The considerable amount of work done in the context of the telephone network can therefore be translated with little effort and adapted in this new environment. The traditional reference framework for the measurement principles for the telephone network is undergoing a major revision process to extend it to other networks: Recommendations E.500 [10], E.502 [12] and E.493 [11] jointly define the set of measurements to be performed by switches for both traffic characterization and performance monitoring.

Additional measurements will be needed as a call in a B-ISDN can be more complicated than a point-to-point connection. A B-ISDN call can consist of multiple ATM connections; also, a given connection can be a point-to-multipoint connection, and in the future multipoint-to-point and multipoint-to-multipoint connections may be defined.

As for traffic characterization, calls are classified according to their bandwidth; for each bandwidth class the quantities of interest are the traffic intensity and its components: the number of bids (call attempts) and the holding times. As the traffic intensity is a continuously varying quantity, simply plotting its time profiles, though giving a qualitative picture of the users' demand of network resources, would not result in sharp indications for the network operator. Hence, the sampled traffic data are subject to averaging on a traffic reference period. It is to be noted that the assumption of one hour as the reference period in telephone networks is consistent both with the typical duration of a telephone call (the reference period is nearly twenty times larger than the mean holding time) and with the time in which the actual call arrival process can be well approximated by a stationary process. In ATM networks, it may be difficult to define a length for the reference period that satisfies both conditions.

The performance (grade of service) at the call level is represented by the blocking probability, and by the delays in the set-up, re-negotiation, and release phases. Of particular relevance in the context of traffic management are the completion ratios, as the Answer-to-seizures ratio and the Answer-to-bids ratio.

3.3. Open issues

1. Is the use of a reference period still a valid paradigm in ATM networks? If so, how should it be defined? As a duration of one hour is assumed for the telephone service (typically twenty times the mean holding time), should its duration be proportional to the mean holding time for the service under consideration? Can we still attain the objective that the reference period is long enough to obtain reliable estimates and yet short enough so that the modeling assumption of stationarity is reasonable?
   1.1. In the presence of long duration calls, the cell-level characteristics may change during the call. Should a reference period be defined for cell level as well as for call level measurements? If so, how can we segment the call duration into blocks each of which can be assumed to exhibit stationarity?
   1.2. The bursty nature of traffic can sometimes be misinterpreted as the lack of stationarity, leading to an overshortening of the reference period. How should we avoid such pitfalls?
2. As the nature of traffic sources gets more varied and the complexity of network mechanisms grows, more complex models are proposed. How to reconcile the desire for accurate traffic characterization with the need to keep the measurement load down to acceptable levels? What parameters should be estimated from measurements?
   2.1. Measurements have revealed that the average value of the cell loss ratio experienced at a switch buffer can vary widely during the call [13]. We would like to have a better understanding of the cell-loss process for given scenarios. For example, for contiguous intervals of time, labeled $i = 1, \ldots,$ each of length T, what is the character of the sequence of cell loss ratios for intervals i,

$i = 1, \ldots,$ for different values of T? How big does T need to be for the variance of the cell loss ratios to be relatively small?

2.2. Is the accurate estimate of performance parameters, such as loss and delay, more difficult if the arrival processes are self-similar? If so, what refinements are needed?

2.3. If self-similarity of arrival processes is pertinent for CAC or dimensioning, what traffic parameters should be estimated?

3. An established way to dimension ATM links is to use the generalized Erlang-B formula. However, this approach may become less useful for the full range of B-ISDN services, including those with elastic traffic characteristics.

3.1. One approach is to adapt the generalized Erlang-B formula by revising the measurements of the input parameters. How should this be done?

3.1.1. In Video-on-Demand, the typical holding time of a connection is of the same order of magnitude as the busy period of connection requests. For example, a popular time to request a movie is between 8:00PM and 9:00PM and these connections are still present during the less popular request period of 9:00PM to 10:00PM, making the latter period have higher occupancy than the former. As a consequence the peak of occupations is shifted in time with respect to the peak of requests [14], so that a steady state is not reached.

3.1.2. Currently, for IP-based applications the point-to-point traffic matrix between network nodes changes much more rapidly, day by day, than for voice traffic. A given web site can be hot for a few days, and then another is.

3.1.3. How should the input parameters be adjusted to account for non-Poisson connection arrivals? Connection requests initiated by humans tend to be Poisson, but not so for those initiated by machines. For example, the requests to initiate an FTP session tend to be Poisson, but no so for the initiations of the file transfers [15].

3.2. Another approach is to use some new method to dimension bandwidth.

3.2.1. For the given method, what measurements would be needed to support it?

4. Long-range dependent and Markovian traffic models

4.1. *Review*

As the concepts of long-range dependence and self similarity have only relatively recently received attention in the teletraffic community, as compared with Markovian models, we begin with a brief summary of key definitions. For additional details, see for example Refs. [16–18].

4.1.1. Long-range dependence and self-similarity

Let $\{X_k, k \geq 1\}$ be a covariance stationary process. Let $r(k)$ be the autocorrelation function $r(k) = E[(X_{1+k} - \mu)(X_1 - \mu)]/\sigma^2$, where $E[X_1] = \mu$ and variance of X_1 is σ^2. If $\sum_{k=0}^{\infty} r(k) = \infty$, then $\{X_k, k \geq 1\}$ is long-range dependent. Likewise, if $\sum_{k=0}^{\infty} r(k) < \infty$ then $\{X_k, k \geq 1\}$ is short-range dependent. If $\{X_k, k \geq 1\}$ is given by a Markovian model, such as a Markov chain or via an underlying process such as the Markov Arrival Process, the autocorrelation function will have a finite sum, and thus the process is short-range dependent. For Markov models generally considered, the autocorrelations eventually decrease exponentially for large k, though if $\{X_k, k \geq 1\}$ is the state of a periodic Markov chain, the autocorrelations can oscillate (and not decay); however the sum remains finite.

A stochastic process $\{Y(t), 0 \leq t < \infty\}$ is self-similar with Hurst parameter H if $\{Y(at)\}$ and $\{a^H Y(t)\}$ have identical finite-dimensional distribution functions for all $a > 0$. Heuristically, for all scalings of time, a self-similar process distributionally looks like itself with the appropriate amplitude scaling. Fractional Brownian motion [19] is a self-similar process. Let $\{X_k, k \geq 1\}$ be an increment process of a self-similar process $\{Y(t), 0 \leq t < \infty\}$: $X_k = Y(k\Delta) - Y((k-1)\Delta)$, where Δ is some fixed time interval. Suppose $\{X_k, k \geq 1\}$ is covariance stationary. Then its autocorrelation function, $r(k)$ has the form $r(k) = (1/2)\delta^2(k^{2H})$, $k \geq 1$, where $\delta^2()$ is the central second difference operator [16], that is:

$$r(k) = \frac{1}{2}\delta^2(k^{2H}) = \frac{1}{2}\left[(k+1)^{2H} - 2k^{2H} + (k-1)^{2H}\right], \quad k \geq 1. \quad (4.1)$$

Note that if $H = 0.5$, then $r(k) = 0$ for $k \geq 1$. If $H \neq 0.5$ then for large k:

$$r(k) \sim H(2H-1)k^{2H-2},$$

where $f(x) \sim g(x)$ means $f(x)/g(x) \to 1$ as $x \to \infty$. Of most interest is the case when $H \in (0.5,1)$, for then $r(k)$ decays as a power law (decays hyperbolically) and the sum of the autocorrelations is infinite. Thus a covariance-stationary increment process of a self-similar process with $H \in (0.5,1)$ is long range dependent.

As an aside, although the self-similar and long-range dependent processes of interest are ‘‘bursty’’, they are not necessarily so. In the context of communication networks consider a constant bit rate connection whose rate is only probabilistically known (such as for dimensioning bandwidth based on a forecast). Consider the trivial model where $Y(t)$ equals the number of bits to arrive over the interval $[0,t]$. Let $Y(t) = Rt$, for some positive random variable R. Let X_k equal the number of bits to arrive in the interval $((k-1)\Delta, k\Delta]$, then $X_k = X_1$, $k \geq 1$. Here $Y(t)$ is self-similar with $H = 1$, and since $r(k)$ equals 1 for all k, $\{X_k\}$ is long-range dependent.

Determining whether a given data set is self-similar is often impractical, as the definition involves the comparison of infinitely many finite dimensional distributions. Cox, in Ref. [17], introduces the practical notion of *second-order self-similarity*, as well as useful, alternative, equivalent characterizations for long-range dependence. We review the key points here.

Again let $\{X_k, k \geq 1\}$ be a covariance stationary process with autocorrelation function $r(k)$, and let $g(\omega)$ be the power spectral density function:

$$g(\omega) = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty} r(k)\,\mathrm{e}^{-\mathrm{i}k\omega}.$$

Define the mth aggregated process, $\{X_k^{(m)}, k \geq 1\}$, $m \geq 1$, via:

$$X_k^{(m)} = m^{-1}(X_{km-m+1} + \ldots + X_{km}), \quad k \geq 1.$$

$\{X_k^{(m)}, k \geq 1\}$ is also covariance stationary. Let ν_m denote the variance$(X_k^{(m)})$, and $r^{(m)}(k)$ denote the autocorrelation function of $\{X_k^{(m)}, k \geq 1\}$. We have the following equivalent conditions for short-range dependence:

1. $\sum_{k=0}^{\infty} r(k) < \infty$,
2. the power spectral density $g(\omega)$ is finite at $\omega = 0$.
3. $\lim_{m\to\infty} m\nu_m < \infty$ (i.e. ν_m is asymptotically of the form: a constant times m^{-1}).
4. $\lim_{m\to\infty} r^{(m)}(k) = 0$, for each $k \geq 1$ (i.e. $\{X_k^{(m)}, k \geq 1\}$ tends to second order pure noise).

Likewise, we have the following equivalent conditions for long-range dependence:

1. $\sum_{k=0}^{\infty} r(k) = \infty$.
2. the power spectral density $g(\omega)$ is singular at $\omega = 0$.
3. $\lim_{m\to\infty} m\nu_m = \infty$.
4. $\lim_{m\to\infty} r^{(m)}(k) \neq 0$, for $k \geq 1$ (i.e. $\{X_k^{(m)}, k \geq 1\}$ has a non-degenerate autocorrelation function).

Cox defines the process $\{X_k, k \geq 1\}$ to be *exactly second-order self-similar* when the aggregated processes not only have a non-degenerate autocorrelation function (and hence are long-range dependent) but have one that is identical to that of the original process:

$$r^{(m)}(k) = r(k), \quad k \geq 1, m \geq 1. \quad (4.2)$$

Moreover, Eq. (4.2) will pertain for the long-range dependent process where $\nu_m = \nu_1 m^{-\alpha}$, $0 < \alpha < 1$. In particular, $r^{(m)}(k)$ will equal Eq. (4.1) above where $H = 1 - \alpha/2$. Lastly, $\{X_k, k \geq 1\}$ is said to be *asymptotically second-order self-similar* when

$$r^{(m)}(k) \to r(k), \text{ as } m \to \infty, \quad k \geq 1,$$

which pertains when $\nu_m \sim cm^{-\alpha}$, $0 < \alpha < 1$, for some constant c. Note that in common usage the ‘‘*second order*’’ is often omitted.

4.1.2. Long-range dependence (LRD) in communication networks

Many recent studies have found that measurements of various data and VBR video traffic exhibit aggregated processes whose variance ν_m decays more slowly than m^{-1}, thus indicating asymptotic (second-order) self-similarity (and long-range dependence). Some particular studies are: Leland et al. [20] in extensive measurement and analysis of Ethernet traffic, Beran et al. [18] in various VBR video sequences of video-conferences, TV programs and movies, and Duffy et al. [21] in common channel signaling traffic in local-exchange-carriers' networks.

It has been noted that when analyzing a finite data set, long-range dependence can be confused with non-stationarity. However, this ambiguity may be an asset as a long-range dependent model may be useful for making a good design decision when the stationarity of the arrival process is in doubt.

Willinger et al. [22] have investigated how long-range dependence of aggregate traffic flows can arise based on characteristics of individual sources. They have shown that if the individual sources are alternating ON/OFF where the length of the ON or OFF period has infinite variance, then the superposition of the sources is long-range dependent. Villén and Gamo [23] show that a process consisting of Poisson arrivals of bursts with infinite variance is long-range dependent. Although no finite data set will have a sample variance of infinity, the data set may be nicely modeled with a distribution with an infinite variance (such as Pareto with a shape parameter less than 2). This has been the case in various measurement studies: for example, for traces of source-destination traffic on a local Ethernet and over a wide-area network [22], for Narrowband-ISDN data traffic [24], for Telnet packet interarrivals and size of FTP bursts [15], for VBR video frame sizes [25], and in world wide web traffic [26].

Given that a process is long-range dependent, a natural follow-on question is whether this attribute is important in the queueing behavior, for regimes of interest. Fowler and Leland [27] study the characteristics of congestion periods using high resolution traces of measured LAN traffic and show a marked difference from that obtained with an analytic source model of Poisson arrivals of fixed size packet trains. Erramilli et al. [28] examine the impact of long-range dependence on queueing behavior of measured Ethernet traffic. They perform experiments by partitioning the data trace into contiguous blocks each consisting of m interarrival times, and then either shuffling the interarrivals within each block (whereby the short-term correlations are eliminated) or shuffling the order of the blocks (whereby the long-term correlations are eliminated) and for both cases the marginal distribution of the interarrival time is unchanged. They find that preserving the order of the blocks (the long-term correlations) is key to matching the true (trace driven) mean delay and complementary queue length distribution.

However, other studies have not confirmed the importance of modeling long-range dependence in order to predict cell loss ratios. Heyman and Lakshman [29] consider traces of VBR video teleconferences and films where the process of number of bits per frame-time is long-range dependent, which they model with a Markov chain (in particular a dynamic autoregressive process of order one), which is short-range dependent. They find that the Markov model accurately estimates the cell loss rates and mean buffer sizes over a wide range of loadings. Similarly, Elwalid et al. [30] show that for video teleconferencing data, Markov chain models can accurately predict the number of connections that can be admitted on a link without violating a cell-loss constraint. Ryu and Elwalid [31] consider model processes based on VBR video where they vary the short term and long term correlations. They find that ''even in the presence of the LRD property, it is the short-term correlations that have a dominant impact on CLRs under realistic scenarios of ATM buffer dimensioning''. Grossglauser and Bolot [32] consider a fluid input to a finite capacity buffer, and they adjust the characteristics of the input rate including its marginal distribution and its autocorrelation function which is hyperbolic up to a cutoff lag. They find that the impact on loss of the correlation in the arrival process becomes nil beyond a given time scale. Also, changes to the marginal distribution (but holding the mean fixed) can significantly influence the loss rate. Andrade and Martínez [33], using the index of dispersion of counts also find that after a given time interval, the correlations in the arrival process have little influence on queue length.

Results by Krishnan [34] may lead to an explanation of some of the above contradictory indications. Consider the queueing of the superposition of independent and identically distributed fractional Brownian motion processes (which are self-similar) where the service rate is adjusted so that the probability the work in system is above a given threshold equals a given value. Consider two cases distinguished by a different value for the Hurst parameter of the sources. Krishnan shows that for sufficiently many multiplexed sources, the required service rate is *smaller* for the sources with the *higher* Hurst parameter [34]. Furthermore, Krishnan and Meempat [35] have shown that this phenomenon also occurs for the VBR video streams used in Ref. [30] where now the sources with the higher Hurst parameter are the traces of the sample video, and the sources with the lower Hurst parameter are from a Markovian model.

4.1.3. Shortcoming of prior work

The above studies which use LRD or Markovian models in the queueing behavior of bursty data applications have considered the arrival processes to be independent of the current state of the network and have not considered their elastic and endogenous characteristic. In reality, the realized flow of, say, IP datagrams associated with a given application depends on the actions of other applications that are also using the network. The flow control and retransmission algorithms of TCP and CSMA/CD influence the resulting flow of datagrams. Measurements of realized Ethernet traffic *do* capture the flow that results from the aggregation of these controls. Researchers have then used these measurements to drive simulations of queueing behavior or to derive analytic models of sources which are then used to drive the simulations or to determine analytic models for the queueing behavior (which can then be computed numerically). However, such studies have treated the arrival process as *independent of* the state of the queue. In reality, if the given process were arriving to a buffer and if losses were occurring, then some control would be altering the future arrivals. For example, in Fowler and Leland's Figs. 11 and 12 [27], the congestion period is 100 seconds wherein losses greater than the objective of 10^{-4} are repeatedly occurring, and wherein losses greater than 10^{-3} occur over a 4 second period. In practice, higher layer controls would have reacted to the losses and reduced the offered traffic; however, as the authors point out, their trace driven simulation ‘‘cannot incorporate the effects of automatic and human responses of congestion’’. In practice, if the Available Bit Rate (ABR) transfer capability is being used, then feedback to selected sources, hopefully prior to cell loss, would alter the future flow of cells. If the Unspecified Bit Rate (UBR) transfer capability is being used, then the discard of a connection's AAL5-PDU would trigger a higher layer control, say TCP, to make the adjustment.

Treating the arrival process as independent of the state of the buffer and imposing a performance objective of low loss is expecting the network to be designed to be transparent for the data applications. In all likelihood, this is economically impractical for public networks. As a point of reference, we should bare in mind that probably *none* of the existing network interface cards, or bridges, routers, or gateways have been designed or are being operated taking into account the possible LRD of the traffic. The endogenous characteristic of the data traffic is one of the reasons why intranets and the Internet work as well as they do.

VBR video also has an elastic quality. In the context of ATM, the VBR real-time transfer capability is tailored to support VBR video. With this transfer capability the cell flow crossing a public interface is suppose to be conforming to the standardized traffic descriptors of PCR and SCR/IBT. In harmony with this requirement, a VBR video encoder, via judicious use of the encoder buffer and adjustments to the quantizer step-size, can be designed to emit cells in conformance with these traffic descriptors, see Ref. [36]. A common worst-case model of such sources is the simple, deterministic ON/OFF source. Another approach for supporting VBR video over ATM is suggested by Kanakia et al. [37] who investigate a dynamic control, like ABR, wherein the encoder responds to feedback from the network – here the resulting cell flow may be LRD but, like the case of data, would be dependent on the congestion state in network nodes.

Thus in the case of VBR video over ATM, if the VBR real-time transfer capability is used, then, although the cell arrival process is independent of the state of the network, it would be conforming to the PCR and SCR/IBT traffic descriptors. If the ABR transfer capability is used, then the cell arrival process would be dependent on the state of the network; studies where the VBR video is modeled as a LRD process that is independent of the state of the network might be relevant for smaller private networks, but as with the case of data, are probably not relevant for public networks.

Some simulation studies have been made with self-similar sources carried over ABR service. Preliminary findings by G Benke et al. [38] are ‘‘that the ABR scheme seems to protect the network well; although larger queues form at the ABR SAP’’. (This study did not seem to include an additional feedback control from the ABR SAP to a higher layer.)

4.2. Open issues

As a rough summary, many important traffic sources exhibit long-range dependence, though Markovian models currently have a richer theory for performance analysis. Evaluation of a given traffic model, whether LRD or Markovian, needs to be done within a stated context. For the ITU’s work on B-ISDN traffic engineering, models of traffic are used to predict performance which in turn are used for dimensioning resources and for connection admission control. The goal is to make ‘‘good’’ decisions, taking into account a vector of concerns. The test of a traffic model is how well it helps in making a good decision.

1. For public ATM networks, we are particularly interested in studies of dimensioning resources and connection admission control for data or video traffic where a control scheme adjusts the source’s emission rate based on congestion in the network.
   1.1. For elastic data applications, where throughput can be a more important performance parameter than connection blocking, how should bandwidth be dimensioned?

5. Conclusion

We have discussed three problem areas of importance to the ITU regarding modeling and measurements for broadband services: (1) traffic descriptors and traffic variables, (2) cell and call-level measurements, and (3) long-range dependent and Markovian traffic models. Further work by teletraffic researchers on these topics and subsequent contributions to the ITU Study Group 2 would be welcomed.

References

[1] ITU-T Recommendation I.371, Traffic Control and Congestion Control in B-ISDN, Geneva, August 1996.

[2] ITU-T Recommendation I.356, B-ISDN ATM Layer Cell Transfer Performance, Geneva, October 1996.

[3] ITU-T Recommendation I.358, Call Processing Performance for Switched Virtual Channel Connections (VCCs) in a B-ISDN, June 1998.

[4] ITU-T Recommendation E.716, User Demand Modeling in Broadband ISDN, Geneva, October 1996.

[5] ITU-T Recommendation E.735, Framework for Traffic Control and Dimensioning in B-ISDN, Geneva, May 1997.

[6] ITU-T Recommendation E.736, Methods for Cell-level Traffic Control in B-ISDN, Geneva, May 1997.

[7] ITU-T Recommendation E.737, Dimensioning Methods for B-ISDN, Geneva, May 1997.

[8] ITU-T Draft Recommendation E.726, Network Grade of Service Parameters and Target Values for B-ISDN, COM2-R40, Geneva, May 1998.

[9] ITU-T Draft Recommendation E.745, Cell Level Measurement Requirements, COM2-R40, Geneva, May 1998.

[10] ITU-T Draft Revised Recommendation E.500, Traffic Intensity Measurement Principles, COM2-R37, Geneva, March 1998.

[11] ITU-T Recommendation E.493, Grade of Service (GoS) Monitoring, Geneva, February 1996.

[12] ITU-T Recommendation E.502, Traffic Measurement Requirements for SPC Telecommunication Exchanges, Geneva, 1992.

[13] R. Grünenfelder, C.J. Gallego, ATM network testing and measurement, Telecommun. Syst. (5) (1996) 241–248.

[14] M. Naldi, Traffic characteristics of multimedia services and their impact on ATM network dimensioning, 13th European Network Planning Workshop ENPW97, Les Arcs, March 1997.

[15] V. Paxson, S. Floyd, Wide-area traffic: the failure of Poisson modeling, IEEE/ACM Trans. Networks 3 (1995) 226–244.

[16] M.S. Taqqu, Self-similar processes, in: S. Kotz, N. Johnson (Eds.), Encyclopedia of Statistical Sciences, vol. 8, Wiley, New York, 1987, pp. 352–357.

[17] D.R. Cox, Long-range dependence: a review, in: H.A. David, H.T. David (Eds.), Statistics: An Appraisal, Iowa State U. Press, Ames, IA, 1984, pp. 55–74.

[18] J. Beran, R. Sherman, M.S. Taqqu, W. Willinger, Long-range dependence in variable-bit-rate video traffic, IEEE Trans. Commun. 43 (1995) 1566–1579.

[19] B.B. Mandelbrot, J.W. Van Ness, Fraction Brownian motions, fractional noises and applications, SIAM Rev. 10 (1968) 422–437.

[20] W.E. Leland, M.S. Taqqu, W. Willinger, D.V. Wilson, On the self-similar nature of ethernet traffic (extended version), IEEE/ACM Trans. Networks 2 (1994) 1–15.

[21] D.E. Duffy, A.A. McIntosh, M. Rosenstein, W. Willinger, Statistical analysis of CCSN/SS7 traffic data from working subnetworks, IEEE J. Select. Areas Commun. 12 (1994) 544–551.

[22] W. Willinger, M.S. Taqqu, R. Sherman, D.V. Wilson, Self-similarity through high-variability: statistical analysis of ethernet LAN traffic at the source level, IEEE/ACM Trans. Networks 5 (1997) 71–86.

[23] M. Villén-Altamirano, J. Gamo, A simple, tentative model for explaining the statistical characteristics of LAN traffic, COM2-R8, ITU Study Group 2, Geneva, 22–31 March 1994.

[24] K. Meier-Hellstern, P.E. Wirth, Y.-L. Yan, D.A. Hoeflin, Traffic models for ISDN data users: office automation application, in: A. Jenson, V.B. Iversen (Eds.), Teletraffic and Datatraffic in a Period of Change, Proc. 13th Int. Teletraffic Congr., Elsevier, Amsterdam, The Netherlands, 1991, pp. 167–172.

[25] M. Garrett, W. Willinger, Analysis, modeling, and generation of self-similar VBR video traffic, in: Proc. SIGCOMM '94, 1994, pp. 269–280.

[26] M.E. Crovella, A. Bestavros, Self-similarity in world wide web traffic: evidence and possible causes, in: Proc. ACM SIGMETRICS, May, 1996.

[27] H. Fowler, W. Leland, Local area network traffic characteristics, with implications for broadband network congestion management, IEEE J. Select. Areas Commun. 9 (1991) 1139–1149.

[28] A. Erramilli, O. Narayan, W. Willinger, Experimental queueing analysis with longe-range dependent packet traffic, IEEE/ACM Trans. Networks 4 (1996) 209–223.

[29] D.P. Heyman, T.V. Lakshman, What are the implications of long-range dependence for VBR-video traffic engineering, IEEE/ACM Trans. Networks 4 (1996) 301–317.

[30] A. Elwalid, D.P. Heyman, T.V. Lakshman, D. Mitra, A. Weiss, Fundamental bounds and approximations for ATM multiplexers with applications to video teleconferencing, IEEE J. Select. Areas Commun. 13 (1995) 1004–1016.

[31] B. Ryu, A. Elwalid, The importance of long-range dependence of VBR video traffic in ATM traffic engineering: myths and realities, in: ACM SIGCOMM '96, pp. 3–14.

[32] M. Grossglauser, J.C. Bolot, On the relevance of long-range dependence in network traffic, in: ACM SIGCOMM '96, pp. 15–24.

[33] J. Andrade, M.J. Martínez-Pascua, Use of the IDC to characterize LAN traffic, 2nd Workshop on Performance Modelling and Evaluation of ATM Networks, IFIP TC 6, Bradford, 4–7 July 1994.

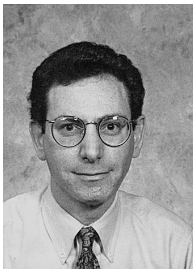
[34] K.R. Krishnan, A new class of performance results for a fractional Brownian traffic model, Queueing Syst. 22 (1996) 277–285.

[35] K.R. Krishnan, G. Meempat, Traffic engineering for VBR video with long-range dependence, in: Proc. Int. IFIP-IEEE Conf. on Broadband Communications, Montreal, 1996, pp. 467–476.

[36] A.R. Reibman, B.G. Haskell, Constraints on variable-bit-rate video for ATM networks, IEEE Trans. Circuits Syst. Video Technol. 2 (1992) 361–372.

[37] H. Kanakia, P.P. Mishra, A.R. Reibman, An adaptive congestion control scheme for real time packet video transport, IEEE/ACM Trans. Networks 3 (1995) 671–682.

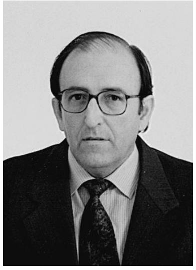
[38] G. Benke, S. Dastangoo, G. Miller, R. Mitchell, N. Schult, M. Procanik, S. Liu, T. Chen, V. Samalan, Simulation results of self-similar traffic over ATM ABR service, ATM Forum, contribution #96-0026, 4–9 February 1996.



Arthur W. Berger received the PhD degree in applied mathematics from Harvard University in 1983. He then joined AT&T Bell Laboratories, and in 1996 AT&T Labs, and for the past twelve years he has been with the Teletraffic and System Analysis Department. In June 1998 he joined Lucent Technologies. He has worked in the areas of network planning, performance analysis of telecommunication switching systems, and in B-ISDN/ATM on the topics of congestion controls and traffic engineering. On the latter topics he has been active in ITU Study Groups 2 and 13, and in the US T1S1 committee and in the ATM Forum. His research interests are in applied probability, congestion controls and traffic engineering for high-speed communication networks. He is a member of IEEE Communications Society and ACM SIGCOMM.

Maurizio Naldi received the Dr. Ing. Degree in Electronic Engineering in 1988 and the Ph.D. in Telecommunications in 1998. From 1989 to 1991 he was with Selenia (now Alenia) as a radar designer. He then joined Italcable, where he has started his involvement in the standardization of broadband networks as a delegate to both the ETSI NA5 Technical Committee and the ITU Study Group 13. In 1995 he has moved to Telecom Italia, where he now works in the Studies Section of the Teletraffic Engineering Department. He is Associate Rapporteur in ITU SG2 for traffic measurements in broadband networks and an external examiner at the University of Rome ''Tor Vergata''. He is a member of AEI, IEEE, and MAA.

Livia De Giovanni received the degree in Applied Statistics in 1988 from the University of Rome ''La Sapienza''. From 1988 to 1997 she worked at Telecom Italia, Network Division Headquarters, Research and Development Department on Broadband Switching. She has worked in the B-ISDN area, with regard to which she has been in charge of producing the specifications of ATM switching systems and related management systems to be used by Telecom Italia. She has also participated, within ITU-T Study Group 2 (Traffic Engineering) standardization body and JAMES (Joint ATM Experiment on European Services) project, to the definition of traffic control functions and traffic measurements reference models in B-ISDN. In March 1997 she joined the University of Molise as a researcher in Probability and Applied Statistics. Her research interests cover statistical inference in queueing systems with application to traffic control functions definition and performance evaluation in ATM telecommunications networks.



Manuel Villén-Altamirano graduated as Master Engineer in Telecommunications from Universidad Politécnica de Madrid in 1970. He then joined ITT (later Alcatel), and in 1989 Telefónica I + D. He has worked in the Traffic Engineering Division of both firms, being head of each of them from 1988 to 1997. He currently occupies the position of senior expert in the Network Analysis and Planning Area. He has participated in several RACE and COST european projects on architecture and performance of ATM and mobile systems and networks. Since 1989 he is Question Rapporteur and from 1995, Vice Chairman of the Traffic Engineering Working Party of the ITU Study Group 2. His research interests are in speed-up simulation techniques, performance and network analysis, traffic controls and dimensioning.